1975

A Method for the Strong Associative Clustering of 2m Data

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Philadelphia: The Annenberg School of Communications, University of Pennsylvania, 1975 (mimeo).

A Method for the Strong Associative Clustering of 2m Data

Draft
Not for Publication or Citation without Permission by the Author
Comment and Criticism will be Appreciated

A METHOD FOR THE STRONG ASSOCIATIVE CLUSTERING OF 2^m DATA

by Klaus Krippendorff

The Annenberg School of Communications
University of Pennsylvania, Philadelphia

February, 1975

Abstract

A METHOD FOR THE STRONG ASSOCIATIVE CLUSTERING OF 2^m DATA

by Klaus Krippendorff
The Annenberg School of Communications
University of Pennsylvania, Philadelphia

A new method of clustering is developed to be applicable on data that consist of a collection of yes and no answers to questions, of checklists of any kind, or of content analysis data of the contingency analysis variety. The method arises from a need for a particular interpretation of the clusters sought which conventional clustering techniques and factor analyses can not warrant. A coefficient for strong association within many-dimensional clusters is proposed. Strong association is contrasted with weak association, and some of the properties of the coefficient are discussed. The algorithm used for strong associative clustering of 2^m data is outlined.

The Problem

Many research designs generate $\underline{2}^{\underline{m}}$ data. For example:

- Interviews consisting of the yes and no answers to $\underline{m}$ preformulated questions
- Experiments in which subjects are asked to push up to $\underline{m}$ different buttons in response to a given stimulus
- The paths of a binary decision tree with $\underline{m}$ branches that a manager may follow in a real or experimental situation.
- Choices among $\underline{m}$ objects whether these are manifest in the form of a shopping list, an expression of preferences for TV shows or whether these indicate knowledge or familiarity with a set of concepts.

The problem that stimulated the research herein reported arises in content analysis, contingency analysis (Osgood, 1959) in particular, in which

- Words, sentences or paragraphs or any convenient linguistic units of analysis are coded each according to whether they contain, represent or imply any of $\underline{m}$ categories or attributes which are then said to be present.

One can represent these data in the form of a collection of m-tuples.

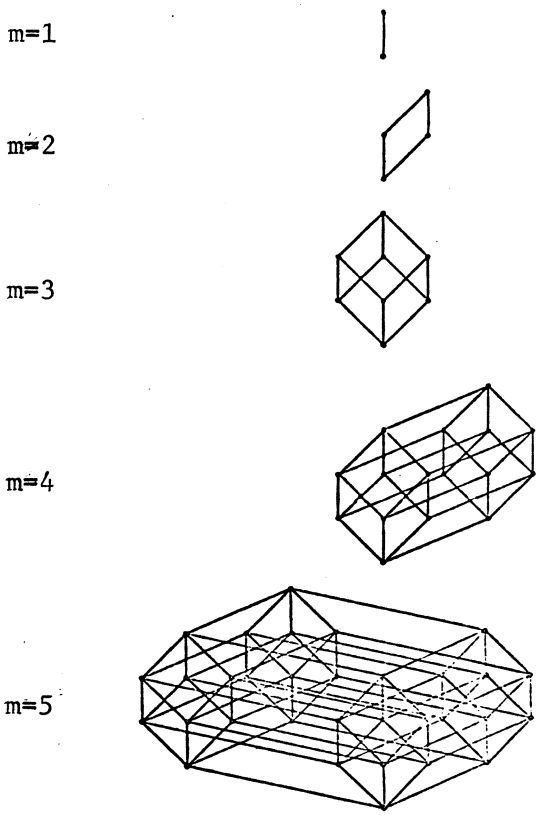
$$
\begin{array}{l}
\langle x_1, x_2, x_3, \cdots, x_m \rangle_1 \\
\langle x_1, x_2, x_3, \cdots, x_m \rangle_2 \\
\langle x_1, x_2, x_3, \cdots, x_m \rangle_3 \\
\quad \vdots \quad \vdots \quad \vdots \quad \ddots \quad \vdots \\
\langle x_1, x_2, x_3, \cdots, x_m \rangle_n
\end{array}
$$

$n \times \underline{2}^{\underline{m}}$ Data

Table 1

With each x_i taking values $\underline{0}$ or $\underline{1}$. And, since there are $\underline{m}$ such dichotomous variables, each datum becomes one of $\underline{2}^{\underline{m}}$ possible ones, hence the name $\underline{2}^{\underline{m}}$ data.

Still another way of looking at the form of these data is that they constitute m-dimensional data cubes. When m is small, problems of visualizing data in those cubes are negligible:



Data Cubes of Different Dimensionality

Figure 1

It is surprising, however, that many researchers cannot conceptualize and, are hence unable to interpret distributions of data in cubes of more than two dimensions which correspond to the classical 2 by 2 table. Others are able to consider a few more dimensions simultaneously but not as many as we wish to analyse. In any of the examples of 2^m data mentioned above, it is not at all uncommon that one has to cope with several hundred dichotomous variables. Without the aid of suitable analytical techniques large numbers of variables make it virtually impossible to render the structure in those data interpretable.

Cluster analysis and factor analysis are two common ways of rendering 2^m data interpretable, the former by collapsing the m-dimensional data cube into one of smaller dimensionality, the latter by superimposing a simple system of coordinates (the factors) on the data in question. Without reviewing these techniques in detail, let us simply say that clusters are formed in such a way that the variables within each cluster are in some sense more similar than variables of different clusters, that they share something that is absent elsewhere, or that they are relatively close to each other, Factors, on the other hand, are chosen so as to represent more efficiently the variation that is in some sense common to a group of variables.

Exactly what is meant by "in some sense similar," or "in some sense common" differentiates among the members of this family of analytical techniques. However, with the advancement of these techniques the notion of "in some sense similar" etc. has become increasingly obscure and results obtained from such analytical techniques have become difficult to interpret in given situations. To avoid getting clustering results that are mere computational artifacts we decided not to rely on any given method - at least not to start out with -- and worked instead on conceptualizing the "sense of similarity" that we felt would yield clusters with interpretations that are both simple and meaningful, at least in the context of our investigative problems.

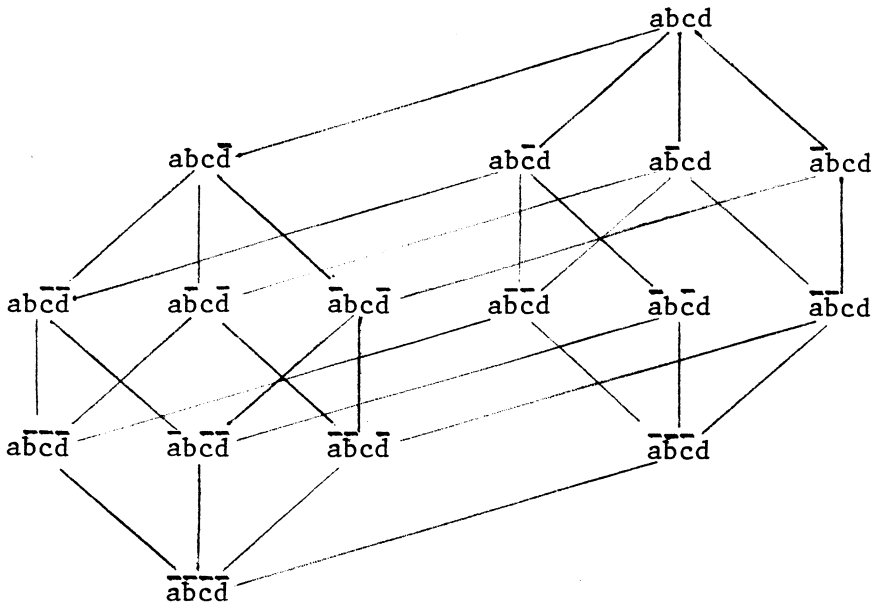
Our approach was initially motivated by the problem of category construction in content analysis. Here, one often wants to develop a system of categories that is least redundant, i.e., one that maps synonymous or semantically similar linguistic units into categories according to what they have in common. Since similarities are not directly representable in 2^m data, the only evidence for "what variables have in common" then lies in the pattern of

cooccurrences. Any clustering of such data must be based on this fact. When the contents or features of linguistic units are described in terms of dichotomous variables, we wanted clustering to be such that all the variables within one cluster become virtually substitutable with a minimum of loss by a single new dichotomous variable, the cluster, which thereby represents exactly what is shared by all of them. Naturally, justifications for "substitutability" and the operationalization of the "loss" involved leads us to statistical consideration of how much variables have in common and how likely this could be the result of chance.

The problem of extracting from data what several variables share is quite a familiar one in social research which leads us to believe that any computational procedure that is developed along these lines will in no way be restricted in its use to content analysis. Our requirements for a strong associative clustering of the variables in 2^m data can be stated quite generally: all resulting clusters should be characterizable by shared patterns of cooccurrences that deviate markedely from chance and from each other.

Strong Association

To consider what variables share, let us depict the possible patterns of cooccurrences that might occur within a four-dimensional data cube:



Possible Cooccurences in $\underline{2^4}$ Data

Figure 2

The quadruple on top of this lattice denotes the pattern of cooccurrences for all four variables, the next level denotes what any three variables share, the next shows which is common to all possible pairs of variables, etc. ending at the bottom in the joint absence of all four variables.

Data assign frequencies to those possible patterns of cooccurrences. For example:

N(abc)

is the frequency with which the variables a, b, and c are observed to cooccur regardless of any presence or absence of variables other than a, b and c. If these three variables are to be considered a cluster then N(abc) denotes the magnitude of what the three variables have in

common while N(ab), N(ac), N(bc) only reflect what its constituent components share. The difference is crucial, as will become clear below.

Frequencies of cooccurrences do not provide a good basis for comparing clusters of different dimensionality because they can only decrease as new variables are added. So, the inequality:

$$N(a) \geq N(ab) \geq N(abc) \geq N(abcd) \tag{1}$$

follows from the fact that all cooccurrences in a cluster of higher dimensionality are contained in each of its constitutent clusters of lower dimensionality.

Furthermore, the frequency of cooccurrences cannot be assumed to vary freely between zero and the sample size N. Unequal frequencies of occurrences impose significant constraints on the frequency range of cooccurrences, particularly in many dimensional data. Again in the three-dimensional case, we have to consider that:

$$0 \leq N(abc)_{min} \leq N(a\bar{b}c) \leq N(abc)_{max} \leq N \tag{2}$$

in which the upper bound of N(abc) is:

$$N(abc)_{max} = \text{Min}\left[N(a), N(b), N(c)\right] \tag{3}$$

and the lower bound of N(abc) is:

$$\begin{aligned} N(abc)_{min} = \text{Max}\Big[& N(a)-N(a\bar{b}c)_{max}-N(\bar{a}bc)_{max}-N(a\bar{b}\bar{c})_{max}, \\ & N(b)-N(a\bar{b}c)_{max}-N(\bar{a}bc)_{max}-N(\bar{a}b\bar{c})_{max}, \\ & N(c)-N(a\bar{b}c)_{max}-N(\bar{a}bc)_{max}-N(\bar{a}\bar{b}c)_{max}, \quad 0 \Big] \end{aligned} \tag{4}$$

Thus, with the frequencies of occurrences N(a), N(b), and N(c) taken as given, the upper bound is obtained whenever all observations that can be

placed in the top and bottom cell of the data cube are indeed found there, whereas the lower bound is obtained whenever these two extremes are avoided as far as possible.

Within the same frequency range lies the frequency that would be expected if cooccurrences were merely due to chance:

$$E(abc) = N^{-2} N(a) N(b) N(c) \tag{5}$$

By extension of (3), (4) and (5) to different numbers of variables it can be shown that the upper and lower frequency bound, as well as the expected frequencies, become smaller as new variables are added:

$$N(a) \geq N(ab)_{max} \geq N(abc)_{max} \geq N(abcd)_{max} \tag{6}$$

$$N(a) \geq E(ab) \geq E(abc) \geq E(abcd) \tag{7}$$

$$N(a) \geq N(ab)_{min} \geq N(abc)_{min} \geq N(abcd)_{min} \tag{8}$$

But they decrease with very different speeds. From the data cubes depicted in Figure 1 it may be seen that -- excepting the two dimensional case -- there are many more ways to stay off the top and bottom of the data cube then to stay on it.

Hence, the upper bound, being affected only by the variable with the smallest number of occurrences, decreases only very slowly, the expected frequency drops with each additional variable, ultimately approximating zero, whereas the lower bound tends to become zero very rapidly. We took this fact as a justification for ignoring the lower frequency bound in the following association coefficient.

A coefficient for strong association that is unbiased by the dimensionality of the data, i.e. by the number of variables in a cluster within which the magnitude of shared cooccurrences is to be assessed is:

$$A(ab...r) = \frac{N(ab...r) - E(ab...r)}{N(ab...r)_{max} - E(ab...r)} \tag{9}$$

Its value is unity when all cooccurrences that could possibly occur do indeed occur within the variables of a cluster and it becomes zero when these cooccurances are merely chance. Negative values are possible too, but, of little interest here.

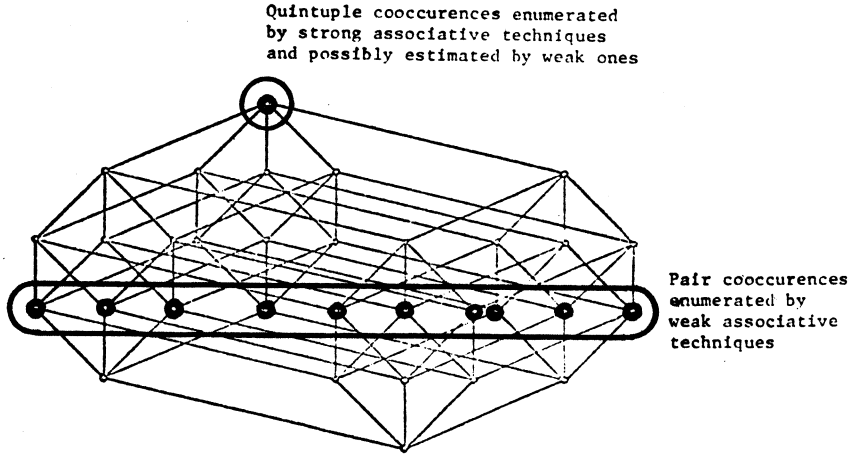
I should like to acknowledge that according to Goodman and Kruskal (1959), the Italian demographer and statistician Benini (1901) used an association coefficient that, like the above, takes the maximum number of cooccurrences into the denominator. However, he was concerned with 2 by 2 tables only. An extension of the notion of association to simultaneous cooccurrences in many variables has to our knowledge not been developed or published to date.

Before discussing some properties of this coefficient an elaboration of our notion of strong association is in order. Aided by the above coefficient we seek to develop clusters that are characterised by a better than chance sharing of cooccurrences in all the variables included in a cluster. Thus, when assessing the association among the three variables of a cluster we enumerate triple cooccurrences and when a fourth variable is added we count quadruple cooccurrences, etc. It is important to notice that what four variables have in common is also shared by the four three-variable clusters and by the six two-variable clusters that are contained in it. This is so because any higher order cooccurrence logically implies the cooccurrence of its lower order components. However, this implication goes one way only.

This fact is also reflected in (1), (3) and (4): given the frequencies of higher

order coocurrences, one can always deduce the smallest number of cooccurrences in its lower order components. But from the frequencies of lower order coccurrences, one can at best deduce the frequency range of higher order cooccurrences. Their actual frequency may be zero or maximum and cannot be induced with certainity.

In contrast to this, all conventional clustering techniques we are aware of, for example, those applied by Sokal and Sneath (1963) those developed by Johnson (1967) and those reviewed by Sibson (1972) including factor analyses start out with and are fundamentally based on correlations, associations, proximities or similarities between pairs of variables. From measures on pairs, each of these techniques have their own ways of estimating what a group of variables "in some sense" share. Associations among the variables within a group are then not of a higher order, rather, they are mere aggregations. Such techniques should be termed weak by comparison. The difference is illustrated in Figure 3 which depicts what strong and weak associative techniques repectively enumerate in order to assess what a five-dimensional cluster shares. With the help of Figure 1, one can easily extrapolate how the discrepancy increases between what a weak technique measures and what it intends to assess as clusters grow in dimensionality.



Cooccurences Enumerated by
Strong and Weak Associative Techniques
Respectively

Figure 3

Now, some of the properties of our coefficient for strong associations (9) remain to be discussed. First of all, strong associations are invariant to the order in which variables are merged to form clusters:

(10) $$A(abc) = A((ab)c) = A(a(bc))$$

This follows from the fact that the cooccurrences N(abc) are enumerated directly from data with component clusters merely pointing to the way as to which are to be paired. The property is important as it assures that results reflect some characteristics of the data and are not biased by the clustering process.

Suppose the data contain a "pure cluster," i.e., a set of variables which are so related that whenever anyone of them is present all others are present and whenever anyone is absent all others are absent. Under these conditions, true associations between all of its pairs, triples, quadruples, etc. are equal and unity in value:

(11) $\{ab...r\}$ is a "pure cluster" $\equiv$ $N = N(ab...r) + N(\bar{a}\bar{b}...\bar{r})$
implies: $A(ab) = ... = A(ar) = A(br) = ... = A(ab...r) = 1$

Suppose the data contain a "truly explanatory variable," say x, all of whose occurrences coincide with the cooccurrences that a set of variables have in common then, the addition of that variable to the set yields an association of unity:

(12) $\{x\}$ is a "truly explanatory variable" $\equiv$ $N(x) = N(ab...rx)$
implies: $A(ab...r) \leq A(ab...rx) = 1$

Conversely, when association within a cluster is unity, the cluster always includes at least one truly explanatory variable. A pure cluster represents the case in which all variables are truly explanatory of each other.

Suppose data contain two associationally equivalent variables or clusters which have the same proportion of cooccurrences within and between these variables, then the association coefficient remains unaffected by the merger:

(13) Two clusters {a...i} and {j...r} are associationally equivalent

$$\equiv N(a...ij...r) = \frac{N(a...i)\ N(j...r)}{N}$$

implies: $A(a...i) = A(j...r) = A(a...ij...r)$

Otherwise, when clusters are merged, or a variable, say s, is added to a cluster, association may increase or decrease. Associations increase whenever coocurrences in the new cluster are relatively more frequent than in its constituent parts:

(14) $$N(a...rs) > \frac{N(s)}{N} N(a...r) \quad \text{or} \quad N(a...rs)_{max} < \frac{N(s)_{max}}{N} N(a...r)_{max}$$

implies: $A(a...rs) > A(a...r)$

And they decrease whenever coocurrences in the new cluster are relatively less frequent than in its constituent components:

(15) $$N(a...rs) < \frac{N(s)}{N} N(a...r) \quad \text{or} \quad N(a...rs)_{max} > \frac{N(s)_{max}}{N} N(a...r)_{max}$$

implies: $A(a...rs) < A(a...r)$

To summarize these properties, (a) the coefficient for assessing higher order cooccurrences in 2^m data is not affected by the order in which clustering proceeds, (b) it is not biased by the dimensionality of the clusters and (c) its value represents the degree to which variables within a cluster can serve as substitutes for the pattern of cooccurring values within that cluster, and (d) the reference to the expected frequencies assures that chance cooccurrences do not enter the measure of strong association.

Statistical Significance of Strong Association

The statistical significance of strong association can be approximated by means of the binominal distribution. Accordingly;

$$(16) \qquad \delta_{\overline{N(ab)>0}} = \sum_{X=N(ab)}^{X=N(ab)_{max}} \frac{N(ab)_{max}!}{X!\,(N(ab)_{max} - X)!} \frac{E(ab)^{X}\,(N - E(ab))^{N(ab)_{max}-X}}{N^{N(ab)_{max}}}$$

This takes account of the upper frequency range of $N(ab)$ but assumes is its lower bound to be zero. The latter may not be justified, however, when $N(a)$ and $N(b)$ are very unequal and the dimensionality of the cluster is very small, e.g. two or perhaps three. But where clustering has reached higher order associations (which is where the power of the method actually lies) this possible bias is likely to have completely disappeared, so that the method can be judged satisfactory for our practical purposes.

The Algorithm

The computation of higher order associations requires repeated references to the original data as depicted in Table 1. We consider these stored in an m by n matrix where m is the number of variables and n is the sample size as has been discussed. The values x_{ij} are either 0 or 1, the 1 denoting the occurence of an attribute or category of that variable. We should like to point out that such a data matrix would not be required in weak associative techniques in which references to data become obsolete once pairwise associations have been obtained.

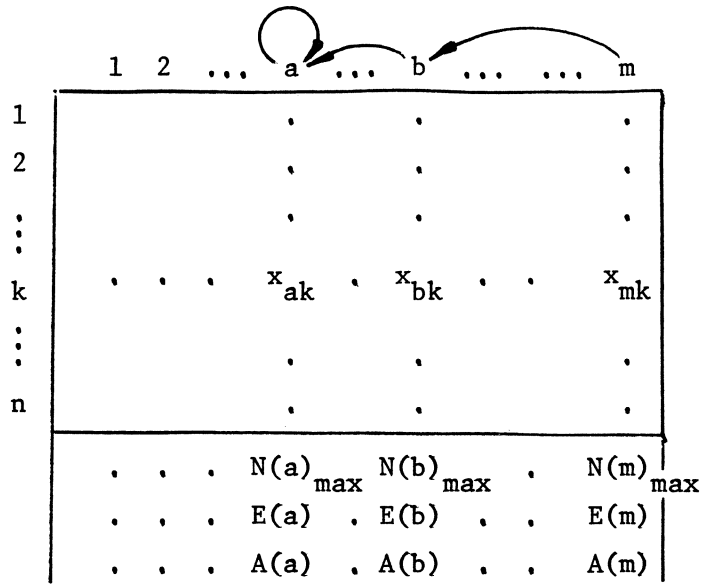
In addition to this data matrix we need two recursively changing values for each variable or cluster, the maximum and the expected frequencies. These are initially set to the observed number of occurences:

$$N(a)^{o}_{max} = N(a)^{o}$$
$$E(a)^{o} = N(a)^{o}$$

to which may be added the association within a cluster, initially set to:

$$A(a)^{o} = 1.$$

These values may be depicted in Table 2 which also illustrates what the merging of variables entails.



Recursively Changing Values

Table 2

We start by enumerating the cooccurences N(ab) within all current pairs of variables or clusters 1 through m by:

$$N(ab) = \sum_{k} x_{ak} \cdot x_{bk} ,$$

compute the association coefficient for each:

$$A(ab) = \frac{N(ab) - N^{-1} E(a) E(b)}{\text{Min}\left[N(a)_{max}, N(b)_{max}\right] - N^{-1} E(a) E(b)}$$

and identify those two variables or clusters for which association is a maximum:

$$a, b \;\Big|\; \text{Max}\left[\ldots, A(ab), \ldots \right]$$

These two variables are then merged onto one by modifying the recursively changing values in Table 2 as follows:

$$N(a)'_{max} = \text{Min}\left[N(a)_{max}, N(b)_{max}\right]$$

$$N(b)'_{max} = N(m)_{max}$$

$$E(a)' = N^{-1} E(a) E(b)$$

$$E(b)' = E(m)$$

$$A(a)' = A(ab)$$

$$A(b)' = A(m)$$

$$x'_{ak} = x_{ak} \cdot x_{bk} \quad \text{for all } k$$

$$x'_{bk} = x_{mk} \quad \text{for all } k$$

$$m' = m - 1$$

After suitably recording the steps for a later account of the hierarchical history of the clustering process, the steps starting with enumerating cooccurences, searching for maximum associations and ending with the modification of the recursively changing values are repeated until some termination criterion is reached.

A nice feature of this algorithm is that the intermediate states of the clustering process can be stored and utilized for restarting the process after inspection, at another point in time and possibly with the addition of more data that would be too costly to process in one run.

The key to the development of this algorithm was a recursive formulation of the maximum and the expected frequencies of occurrencies in many-dimensional clusters. In addition, these had to be invariant to the order in which variables are merged into clusters, one by one. This part turned out to

be very efficient computationally. However, the repeated references that are required to obtain cooccurrences from the data matrix is costly both in time as well as in the storage space required. Strong associative clustering is therefore more limited in scope than many of the weak associative techniques. But this is the price that needs to be paid for results which are more easily interpretable. Nevertheless, a computer program that implements this algorithm has already been applied to a 300 variable - 800 case sample though at not so negligible costs. The future may bring improvements in efficiency of this algorithm.

References

Benini, Rodolfo, Principii di Demografia, Firenzi: G. Barbera, 1901. No. 29 of Manuali Barbera di Scienze Giurichiche Sociale e Politiche.

Goodman, Leo A. and William H. Kruskal, "Measures of Associations for Cross Classification. II: Further Discussions and References. Journal of the American Statistical Association 54: 123-163, March 1959.

Johnson, Steven C., "Hierarchical Clustering Schemes, "Psychometrica 32: 241-254, 1967.

Osgood, Charles E. "The Representational Model and Relevant Research Methods," pp 33-88 in Ithiel de Sola Pool (Ed.) Trends in Content Analysis, Urbana: University of Illinois Press, 1959.

Sibson, Robin, "Order Invariance Methods for Data Analysis," Journal of the Royal Statistical Society 3: 311-349, 1972.

Sokal, Robert R. and Peter H. A. Sneath, Principles of Numerical Taxonomy, San Francisco: W. H. Freeman, 1963.

List of Figures and Tables